Int. J. Appl. Math. Comput. Sci., 2023, Vol. 33, No. 3, 419–428
DOI: 10.34768/amcs-2023-0030

SEMI–SUPERVISED VS. SUPERVISED LEARNING FOR MENTAL HEALTH MONITORING: A CASE STUDY ON BIPOLAR DISORDER

GABRIELLA CASALINO[a,*], GIOVANNA CASTELLANO[a], OLGIERD HRYNIEWICZ[b],
DANIEL LEITE[c], KAROL OPARA[b], WERONIKA RADZISZEWSKA[b],
KATARZYNA KACZMAREK-MAJER[b]

[a] Department of Computer Science
University of Bari Aldo Moro
Via E. Orabona 4, Bari 70125, Italy
e-mail: gabriella.casalino@uniba.it

[b] Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
e-mail: k.kaczmarek@ibspan.waw.pl

[c] Department of Engineering and Science
Adolfo Ibanez University
Pdte Errazuriz 3485, 7550344 Santiago, Chile

Acoustic features of speech are promising as objective markers for mental health monitoring. Specialized smartphone apps can gather such acoustic data without disrupting the daily activities of patients. Nonetheless, the psychiatric assessment of the patient's mental state is typically a sporadic occurrence that takes place every few months. Consequently, only a slight fraction of the acoustic data is labeled and applicable for supervised learning. The majority of the related work on mental health monitoring limits the considerations only to labeled data using a predefined ground-truth period. On the other hand, semi-supervised methods make it possible to utilize the entire dataset, exploiting the regularities in the unlabeled portion of the data to improve the predictive power of a model. To assess the applicability of semi-supervised learning approaches, we discuss selected state-of-the-art semi-supervised classifiers, namely, label spreading, label propagation, a semi-supervised support vector machine, and the self training classifier. We use real-world data obtained from a bipolar disorder patient to compare the performance of the different methods with that of baseline supervised learning methods. The experiment shows that semi-supervised learning algorithms can outperform supervised algorithms in predicting bipolar disorder episodes.

Keywords: semi-supervised learning, mental health monitoring, acoustic features, pattern recognition, AI in medicine.

1. Introduction

Acoustic parameters extracted from speech have been recently studied as objective biomarkers for a psychiatric assessment of the mental state to complement otherwise invasive and costly methods (Arevian *et al.*, 2020; Panek *et al.*, 2015). They are particularly promising for predicting a recurrence of bipolar disorder (BD) episodes (Faurholt-Jepsen *et al.*, 2019). The presence of smartphones in everyday life facilitates the continuous and nondisruptive collection of acoustic data. At the same time, the main challenge of sensor-based mental health monitoring remains in proper data analysis (Antosik-Wójcińska *et al.*, 2020). While acoustic data can be collected during virtually every phone call, usually the assessment of the mental state of a patient seldom occurs. If it occurs, it typically happens during onsite interviews with psychiatrists. Moreover, the process of assigning labels is accompanied by several uncertainties, i.e., when annotating the audio recordings with class labels. Furthermore, the outcome of a psychiatric interview is subject to the patient's condition during the visit. Nonetheless, the patient's mental state may change radically after the visit, especially in the case of treatment modification. Researchers typically extrapolate the BD

[*] Corresponding author

phase of a patient (as assessed by a psychiatrist) to the surrounding days assuming some specific ground truth for the analyses, such as 7 days before and 2 days after the psychiatric assessment (Grünerbl *et al*., 2015).

Within this paper, acoustic features extracted from the speech are preprocessed and semi-supervised learning is applied to capture uncertainties related to partially labeled data collected from smartphones. This study is a continuation of our previous works concerning smartphone-based monitoring of speech. The majority of the related work in this application domain limits the considerations only to labeled data using a predefined ground-truth period (see, e.g., Grünerbl *et al*., 2015; Espinola *et al*., 2021; Dominiak *et al*., 2022). Previously (Casalino *et al*., 2020), we have proposed the use of an incremental semi-supervised classification algorithm based on fuzzy C-means clustering algorithm. However, the lack of labels was simulated. This algorithm was further extended by Kaczmarek-Majer *et al.* (2022b; 2022c) to reflect the dynamic nature of data in the definition of linguistic summaries constructed for explaining purposes, and by Kmita *et al.* (2022) to handle label uncertainty. To ensure proper aggregation of sensor data, the mental changes are assigned to day periods rather than individual acoustic frames in data (Hryniewicz and Kaczmarek-Majer, 2021; Kamińska *et al*., 2020). In the present work, we take one step further in this direction and perform an experimental evaluation of top-performing semi-supervised and supervised classifiers. While the proposed approach makes use of previously existing methods, its main novelty is performing common evaluations for real-life data (both labeled and unlabeled data are considered).

The major contribution of this work is confirming that semi-supervised learning outperforms supervised learning for partially labeled data streams in the context of mental health monitoring. The performance of the semi-supervised methods has been illustrated with data from a prospective study carried out by the Institute of Psychiatry and Neurology in Warsaw (Poland), and the applied methods were evaluated in terms of accuracy using cross-validation and out-of-time scenarios.

That paper is organized as follows. In Section 2, related work is presented. Section 3 focuses on the methodology. Section 4 explains the application scenario. The results of the experiments are presented in Section 5. The conclusion and future directions of this research are depicted in Section 6.

2. Literature review

In the last two decades, machine learning and its applications in healthcare have gained a lot of attention. Learning algorithms are used to support medical decisions by automating time-consuming tasks (Alanazi, 2022; Kusy and Zajdel, 2021). Among the machine learning algorithms, supervised techniques require labeled data, and providing labels is usually a very tedious and error-prone task. In many applied contexts, collecting labeled data is even infeasible, or only very limited labeled examples can be gathered. At the same time, unsupervised learning techniques, such as clustering algorithms, overcome these limitations since they extract knowledge from unlabeled data to construct predictive models. However, due to the absence of labeled information on the data distribution, clustering methods may provide non-satisfactory results and generate data partitions that include instances from different *a-priori* known classes.

Semi-supervised learning algorithms aim at using a combination of both labeled and unlabeled data. We follow the key idea of semi-supervised learning to equip unsupervised learning (e.g., clustering) with a partial supervision mechanism that provides useful guidelines during the process of knowledge discovery from data. Hence, the goal of semi-supervised clustering is to identify semantic categories by grouping together similar data taking advantage of the domain knowledge (supervised information) explicitly supplied by a domain expert (González-Almagro *et al*., 2023).

The use of class labels to aid unsupervised clustering has been the focus of several research works and various computational methods have been applied in the context of clustering with partial supervision ranging from seed, model, and expectation-maximization, support vector machines (Li *et al*., 2009b), probabilistic methods (Ao *et al*., 2014; Ruiz and Finke, 2019), to objective function based algorithms (Bilenko *et al*., 2004) such as K-means (Basu *et al*., 2002).

In the last few years, many researchers have proposed semi-supervised fuzzy clustering algorithms. In particular, a huge number of semi-supervised variants of the well-known FCM (fuzzy C-means) clustering algorithm (Pedrycz and Waletzky, 1997) has been proposed (see, e.g., Li *et al*., 2009a; Mai and Ngo, 2015; Arshad *et al*., 2019; Lai and Garibaldi, 2011). Fuzzy C-Means (FCM) by Bezdek (2013) allows the data to be allocated to several clusters (classes) to various degrees. In the work of Arshad *et al.* (2019), a semi-supervised fuzzy C-means clustering for imbalanced multi-class classification is proposed. In the work of Bennett and Demiriz (1998) a variant of the support vector machine (SVM) is proposed for semi-supervised data. Affinity graphs are used by Zhu and Ghahramani (2002) as well as Zhou *et al.* (2003) to iteratively propagate the class labels from labeled data to their unlabeled neighbors. In the work of Yarowsky (1995) an algorithm to use unlabeled data with a supervised algorithm is presented. For a complete review of semi-supervised methods, the reader is referred to Cai *et al.* (2023).

Notwithstanding the number of semi-supervised algorithms proposed in the literature, to the best of our knowledge, there has been no systematic attempt yet to investigate the efficiency of these algorithms in sensor-based mental health monitoring, which is highly characterized by the presence of both labeled and unlabeled data of uncertain nature. In the majority of related works (see, e.g., Low *et al*., 2020; Espinola *et al*., 2021; Arevian *et al*., 2020; Grünerbl *et al*., 2015; Faurholt-Jepsen *et al*., 2019), the problem of predicting a new bipolar disorder episode is stated as a supervised learning task. To alleviate the challenges deriving from the uncertainty about patients' state and limited data, Hryniewicz and Kaczmarek-Majer (2021) as well as Kamińska *et al*. (2019) applied unsupervised learning techniques are applied and the whole dataset was used for learning rather than constraining it only to a few days before and after the interview with the psychiatrist.

3. Methodology

The aim of this work is to compare selected supervised and semi-supervised methods in the context of sensor-based monitoring of mental state.

3.1. Supervised learning. The classification task consists of finding a function $f : \mathbb{R}^p \rightarrow C$ that assigns one of k labels $C = \{c_1, \ldots, c_k\}$ to each observation vector $x_j \in \mathbb{R}^p$. In our case, this translates to indicating the likely mental state of a patient based on a given set of acoustic features of her speech. The training of classifiers is based on a list of n observations $X = [x_1, \ldots, x_n]$. Supervised methods require it to be fully labeled, i.e., to provide a vector of classes $Y = [y_1, \ldots, y_n]$, where $y_i \in C$. In experiments, we apply decision trees (DTs), K-nearest neighbors (KNN), and the support vector machine (SVM) as baseline methods. For further details, see, e.g., the work of Breiman *et al*. (2017). According to a recent review by Antosik-Wójcińska *et al*. (2020), these supervised algorithms appear frequently in the literature to support the considered applied problem.

3.2. Semi-supervised learning. Let us now consider a scenario that for l ($l < n$) observations from X the class label is unknown, thus it is a partially labeled dataset. We formulate the task of semi-supervised learning as finding a function $f : \mathbb{R}^p \rightarrow C^*$ that assigns one of $k + 1$ labels $C^* = \{c_1, \ldots, c_k\} \cup \{\text{NA}\}$ to each observation vector $x_j \in \mathbb{R}^p$. Keeping the unlabeled observations allows semi-supervised learning algorithms to exploit their patterns to improve prediction. This is most useful for scarcely-labeled datasets. Alternatively, omitting the observations x_j that lack the class assignment, $y_j = \text{NA}$, from the training data would reduce the problem to the supervised case.

Semi-supervised algorithms are usually based on two main assumptions: (i) *assumption of consistency*: similar observations are more likely to belong to the same class. This is a local assumption; (ii) *cluster assumption*: data belonging to the same geometrical structure (e.g., clusters) are more likely to share the same label. It is a global assumption (Zhou *et al*., 2003). In this work, we compare the following four semi-supervised classification algorithms: the semi-supervised support vector machine (S3VM), label spreading (LS), label propagation (LP), and the self training classifier (STC). They are now briefly presented.

3.2.1. Semi-supervised support vector machine. The supervised support vector machine (SVM) algorithm estimates the classification function by using the principle of statistical risk minimization (SRM). However, this function takes into account labeled data only. On the contrary, overall risk minimization (ORM) is able to learn a classification function by minimizing both the empirical misclassification rate and the function capacity on labeled and unlabeled data. The semi-supervised support vector machine (S3VM) algorithm[1] was originally proposed by Bennett and Demiriz (1998). It learns the model from a "training set" (labeled data) and a "working set" (unlabeled data) by constructing a support vector machine to solve the ORM problem proposed by Vapnik (2006).

3.2.2. Label propagation. The label propagation (LP) algorithm, proposed by Zhu and Ghahramani (2002) is an iterative algorithm that propagates labels through the data using a high-density area obtained from the unlabeled data. This algorithm exploits the assumption of consistency: observations close to each other share similar labels. Thus, a fully connected graph is generated, where nodes are data points and the edges are weighted based on the Euclidean distance. Labels are then propagated from each node to its neighbors, according to the distance. Particularly, soft labels, interpreted as distributions over labels, are generated from each node to all the connections.

3.2.3. Label spreading. The label spreading (LS) algorithm[2], proposed by Zhou *et al*. (2003), is a variant of the LP algorithm. It uses a symmetrical matrix to spread the labels through a fully connected graph. During each iteration, each node receives information from its neighbors and at the same time preserves the initial information. Labels are assigned to unlabeled data

[1] S3VM Python library: `https://pypi.org/project/semisupervised/`.

[2] LS, LP and STC are available in the Scikit-learn semi-supervised learning library: `https://scikit-learn.org/stable/modules/semi_supervised.html`.

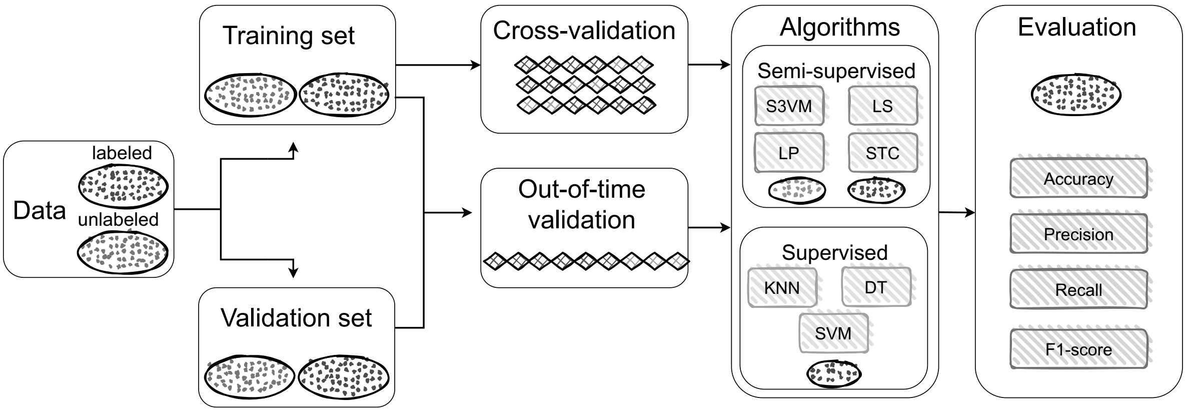


Fig. 1. Workflow explaining the experimental setting.

based on the majority of information received during the iterative process.

3.2.4. Self training classifier. The self-training classifier (STC) proposed by Yarowsky (1995) allows using supervised classifiers as semi-supervised, in order to learn from unlabeled data. It is an iterative algorithm that predicts pseudo-labels for the unlabeled data and adds them to the training set. The algorithm continues iterating until a stop condition is reached (e.g., maximum number of iterations or no new pseudo-labels are added). In this work, we adapt the support vector machine with the radial basis function (RBF) kernel.

4. Application scenario: Classification of bipolar disorder episodes

4.1. About the prospective BD study. Bipolar disorder (BD) is a chronic disease affecting 1–2% of the population (Grande *et al.*, 2016). It is characterized by various episodes ranging from euthymia (state of health) to the mixed states (depressive and manic symptoms present). Early detection of a starting episode is important for improved treatment, however, the frequency of visits with the psychiatrist is usually insufficient to provide early intervention, and patients by themselves are usually not aware of the need for treatment if a new episode starts. Hopefully, smartphones may deliver valid objective markers, such as acoustic features extracted from speech (Antosik-Wójcińska *et al.*, 2020). A recent prospective study was conducted in 2017 and 2018 in the Department of Affective Disorders, Institute of Psychiatry and Neurology in Warsaw (see the work of Dominiak *et al.* (2022) for the protocol of this study) to further investigate these markers. The study included patients diagnosed with bipolar disorder (according to ICD-10 classification). BDmon—a dedicated mobile application—was developed and installed on the patients' smartphones to collect acoustic features from the patient's voice during phone calls.

4.2. Acoustic features extraction and data pre-processing. The acoustic parameters were extracted from voice data using the openSMILE library (Eyben *et al.*, 2013), calculated for the short frames of 20 ms and omitting the interlocutor's speech. The variability of the following acoustic parameters was considered as predictors in further analyses: (i) energy of the speech signal; (ii) the fundamental frequency (F0) and its envelope, which is the dominant tone; (iii) zero-crossing rate at which a signal changes its sign; (iv) voicing probability; and (v) the mel-frequency cepstral coefficients (MFCC). Next, noise or silence was removed from the signal with methods inspired by Otsu (1979) and the levels of the loudness parameter. Finally, each call was aggregated using arithmetic mean and standard deviation.

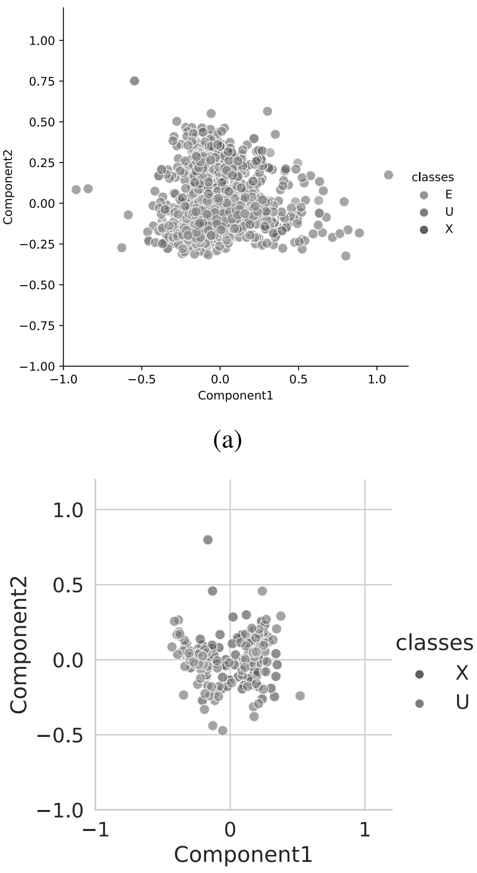
4.3. Labeling. We now explain the ground truth and the labeling process based on psychiatric assessments. During each interview, the doctor assessed the patient's mental state using questions concerning the depressive symptoms derived from the Hamilton depression rating scale (HAMD) and the manic symptoms derived from the young mania rating scale (YMRS). The higher the total score, the more intense the depressive or manic symptoms are. Next, based on the outcomes of the psychiatric assessment and the ground-truth period following (Grünerbl *et al.*, 2015; Faurholt-Jepsen *et al.*, 2016), each observation was assigned a healthy class (euthymia denoted as E), unhealthy class (mixed episode denoted as X) or no label was assigned (denoted as U). No depressive (D) or manic (M) classes were present for the patient, whose data is the basis of this study.

4.4. Experimental setting. Figure 1 summarizes the adopted experimental setting. We address the bipolar disorder prediction as a binary classification problem where we predict the healthy and the unhealthy episodes. First, we consider a subset of data (training set) to compare the different classification algorithms. Stratified cross-validation is used in order to obtain general results. For a fair comparison, the same subsets of data are considered for each fold and only labeled data are used for the supervised algorithms, whilst both labeled and unlabeled data are used for the semi-supervised algorithms. Standard classification measures have been used to evaluate the classification performance for both supervised and semi-supervised algorithms; thus, only labeled data, in each fold, are used for the performance evaluation.

However, these global models, obtained through cross-validation, do not capture the characteristics of data that evolve over time. In order to overcome this limit, we validated the approach in a realistic setting, and out-of-time validation was performed. Phone calls surrounding the last visit with the psychiatrist during which the state of the patient was confirmed and all the proceeding phone calls are considered as validation data. Classification algorithms were trained on the training set and then evaluated on the validation set. In this way, even if still in a static setting, time has been taken into consideration. In the context of bipolar data, the analysis cannot ignore when data were acquired, because predictive models should be able to identify the bipolar episodes in order to alert in the case of an onset of a disease state.

5. Experimental results

5.1. About data. In this work, data belonging to a single patient are considered. A total of 1035 calls were split into two sets (training and validation). As summarized in Table 1, the patient was in the healthy state (euthymia—E) and in the unhealthy state (mixed—X), but most of the frames are unlabeled (U). Indeed, since the control visits cover a low number of days, several unlabeled data have been collected before and after when these states were recognized. The sequences of the states, for the training and validation sets, respectively, were the following: “UXUEU” and “XU”. In the calls considered for the training set, the patient moved from the disease state to the healthy one. The validation set contains only disease data and unlabeled ones. Figure 2 shows the data distribution for the training and validation sets. Principal component analysis (PCA) has been used to reduce the data dimensions. It can be observed that (i) the healthy and disease classes are mostly overlapped (E and X dots), (ii) unlabeled data points (U dots) are the majority, and (iii) data distribution in training and validation sets are very different.



(a)

(b)

Fig. 2. Visualization of training (a) and validation (b) datasets in a two-dimensional space, obtained through PCA.

Table 1. Main characteristics of datasets.

	State	No. of calls	Total
Training set	X	122	863
	E	66	
	U	675	
Validation set	X	80	172
	U	92	

5.2. Cross-validation. First, we compare the performance using the standard cross-validation setting on data belonging to the “training set.” Five-fold cross-validation was used to evaluate the robustness of the classification algorithms. Table 2 shows the classification performance of the considered algorithms. Particularly, the average values, through the folds, are reported. We can observe that in this setting the best results were returned by the S3VM algorithm (semi-supervised), which overcomes the supervised ones. It returns high precision and recall values, suggesting that the model is able to correctly recognize the healthy and disease

Table 2. Average values of classification performance, obtained with cross-validation setting, on the training data.

Algorithm	Accuracy	Precision	Recall	F1-score
KNN	0.79	0.80	0.74	0.75
DT	0.70	0.67	0.67	0.67
SVM	0.80	0.80	0.74	0.75
LS	0.70	0.82	0.56	0.54
LP	0.73	0.83	0.62	0.60
STC	0.79	0.80	0.74	0.75
S3VM	**0.89**	**0.90**	**0.87**	**0.88**

Table 3. Classification performance with out-of-time validation setting on the validation set.

Algorithm	Accuracy	Precision	Recall	F1-score
KNN	0.81	1	0.81	0.90
DT	0.60	1	0.60	0.75
SVM	0.92	1	0.93	0.96
LS	**0.96**	**1**	**0.96**	**0.98**
LP	**0.96**	**1**	**0.96**	**0.98**
STC	0.71	1	0.71	0.83
S3VM	0.80	1	0.80	0.89

episodes. This result suggests that the algorithm is able to exploit the hidden information in data to create a model that better fits the data if compared with the supervised algorithms that could only use *a-priori* information. Also, the label spreading and label propagation algorithms return comparable results with the decision tree, but the KNN and SVM perform better, whilst, the self-training classifier has comparable results with the KNN and SVM.

We assembled violin plots to illustrate the stability of the considered methods. This visualization method simultaneously presents a box plot and kernel density estimate. Panels in Fig. 3 show the following performance metrics: (a) accuracy, (b) precision, (c) recall, (d) F1-score. Compact violins indicate a higher repeatability of the results than stretched ones.

The decision tree algorithm is the most stable, which is indicated by its compact and wide violins for the four measures, suggesting that it returns similar values irrespective of the randomness related to the fold assignment in cross-validation. However, as previously discussed, its classification performance is low. The S3VM algorithm archives the best classification values, and it is quite stable. It is worth noticing that the KNN and SVM algorithms are able to reach quite high classification values, but their plots are stretched, suggesting sensitivity to data. Among the semi-supervised methods, LS, LP, and STC have very thin and long violins for the F1-score graph, suggesting not stable results. However, bipolar data are very difficult to classify, as confirmed by previous works. Also, the assumption of consistency and the cluster assumption is hardly satisfied in bipolar data as shown in Fig. 2 classes are not easily separable, and they are overlapped. Also, no patterns can be identified in data. All those factors affect the outcomes.

5.3. Out-of-time validation. Table 3 summarizes the results obtained with the train-test setting. In this case, the LS and LP algorithms outperform the fully supervised ones. Even though all labeled data in the validation set belongs to the disease class, DT and STC models have the lowest recall values, suggesting the presence of false negatives. On the contrary, LS and LP reach very high values of recall (almost 1), thus all the observations were correctly assigned to the disease class.

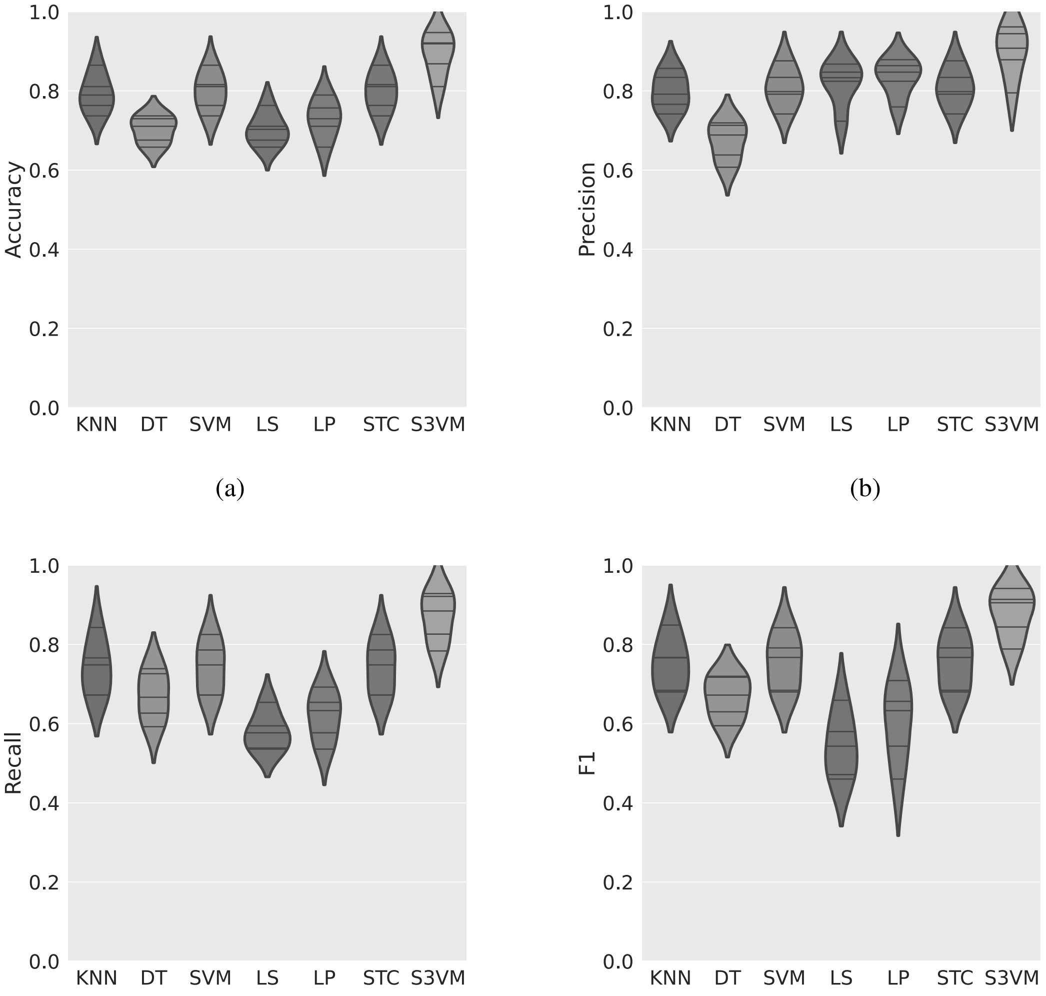
Overall, in both cross-validation and train-test settings, we observe that the semi-supervised algorithms return better results than the supervised ones. In the context of bipolar data, where most of the data are unlabeled, this is an encouraging result, suggesting that both labeled and unlabeled data should be considered for more accurate predictions.

6. Conclusion

Semi-supervised learning algorithms have been increasingly gaining attention in recent years due to the extensive presence of data with partial labeling. We experimentally illustrated that semi-supervised algorithms can outperform supervised ones for a mental health monitoring problem. Semi-supervised methods are able to exploit both the *a-priori* knowledge coming with the class labels, and the inner structure of data, emerging in an unsupervised way.

In this work, we consider a case study on the classification of bipolar disorder episodes for a single patient. These are stream data representing acoustic features derived from frames of patients' calls, collected with a dedicated mobile application. Whilst acoustic features are obtained on a daily basis, labels, corresponding to patients' states, are assigned during control visits occurring every 2–3 months. Hence, the data are naturally only partially labeled. Two validation scenarios have been carried out to compare the classification performance of semi-supervised and supervised algorithms. First, we run stratified cross-validation to evaluate the robustness and generality of the selected methods, and we then run out-of-time validation. In both scenarios, semi-supervised algorithms outperformed the supervised ones. Results have shown that semi-supervised algorithms play an important role in discovering hidden structures in data, so as in improving the classification performance in the presence of a limited fraction of labeled data.

In future work, we plan to conduct experiments



(a)

(b)

(c)

(d)

Fig. 3. Comparison of different algorithms in terms of classification accuracy (a), precision (b), recall (c), and F1-score (d).

for a larger and more diverse group of patients. Methodologically, we intend to exploit fuzzy semi-supervised clustering as a knowledge-based guidance mechanism to provide additional hints about data. The evolving nature of data will be exploited through stream semi-supervised algorithms (Gomes *et al.*, 2022; Leite *et al.*, 2020). Finally, multimodal data, combining different kinds of information will be used to verify whether they are able to improve both the classification performance and the explainability of the results (Kaczmarek-Majer *et al.*, 2022a).

Acknowledgment

Katarzyna Kaczmarek-Majer received funding from the Small Grants Scheme (NOR/SGS/BIPOLAR/-0239/2020-00) within the research project *Bipolar Disorder Prediction with Sensor-based Semi-supervised Learning* (*BIPOLAR*).

Datasets considered in this paper were collected in the CHAD project entitled *Smartphone-based Diagnostics of Phase Changes in the Course of Bipolar Disorder* (RPMA.01.02.00-14-5706/16-00) financed from EU funds (Regional Operational Program for Mazovia) in 2017 and 2018. Gabriella Casalino acknowledges funding from the European Union PON project *Ricerca e Innovazione*, 2014–2020, DM 1062/2021.

Gabriella Casalino and Giovanna Castellano are members of the INDAM GNCS research group. The authors thank psychiatrists and patients who participated in the observational study for their commitment. The authors thank the researchers Monika Dominiak, Anna Wójcińska and Łukasz Święcicki from the Institute of Psychiatry and Neurology for their advice and comments.

References

Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities, *Informatics in Medicine Unlocked* **30**: 100924.

Antosik-Wójcińska, A.Z., Dominiak, M., Chojnacka, M., Kaczmarek-Majer, K., Opara, K.R., Radziszewska, W., Olwert, A. and Łukasz Świecicki (2020). Smartphone as a monitoring tool for bipolar disorder: A systematic review including data analysis, machine learning algorithms and predictive modelling, *International Journal of Medical Informatics* **138**: 104131.

Ao, X., Luo, P., Ma, X., Zhuang, F., He, Q., Shi, Z. and Shen, Z. (2014). Combining supervised and unsupervised models via unconstrained probabilistic embedding, *Information Sciences* **257**: 101–114.

Arevian, A.C., Bone, D., Malandrakis, N., Martinez, V.R., Wells, K.B., Miklowitz, D.J. and Narayanan, S. (2020). Clinical state tracking in serious mental illness through computational analysis of speech, *PLoS ONE* **15**(1): e0225695.

Arshad, A., Riaz, S. and Jiao, L. (2019). Semi-supervised deep fuzzy c-mean clustering for imbalanced multi-class classification, *IEEE Access* **7**: 28100–28112.

Basu, S., Banerjee, A. and Mooney, R. (2002). Semi-supervised clustering by seeding, *Proceedings of the 19th International Conference on Machine Learning (ICML-2002), Sydney, Australia*.

Bennett, K. and Demiriz, A. (1998). Semi-supervised support vector machines, *in* M. Kearns *et al.* (Eds), *Advances in Neural Information Processing Systems*, Vol. 11, MIT Press, Cambridge, pp. 368–374.

Bezdek, J.C. (2013). *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

Bilenko, M., Basu, S. and Mooney, R.J. (2004). Integrating constraints and metric learning in semi-supervised clustering, *Proceedings of the 21st International Conference on Machine Learning, Banff, Canada*, p. 11.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (2017). *Classification and Regression Trees*, Routledge, New York.

Cai, J., Hao, J., Yang, H., Zhao, X. and Yang, Y. (2023). A review on semi-supervised clustering, *Information Sciences* **632**: 164–200.

Casalino, G., Castellano, G., Galetta, F. and Kaczmarek-Majer, K. (2020). Dynamic incremental semi-supervised fuzzy clustering for bipolar disorder episode prediction, *in* A. Appice *et al.* (Eds), *Discovery Science, DS 2020*, Lecture Notes in Computer Science, Vol. 12323, Springer, Cham, pp. 79–93.

Dominiak, M., Kaczmarek-Majer, K., Antosik-Wojcinska, A.Z., Opara, K.R., Wojnar, M., Olwert, A., Radziszewska, W., Hryniewicz, O., Swiecicki, L. and Mierzejewski, P. (2022). Behavioural data collected from smartphones in the assessment of depressive and manic symptoms for bipolar disorder patients: Prospective observational study, *Journal of Medical Internet Research* **24**(1): e28647.

Espinola, C.W., Gomes, J.C., Pereira, J.M.S. and dos Santos, W.P. (2021). Detection of major depressive disorder using vocal acoustic analysis and machine learning—An exploratory study, *Research on Biomedical Engineering* **37**: 53–64.

Eyben, F., Weninger, F., Gross, F. and Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor, *Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain*, pp. 835–838.

Faurholt-Jepsen, M., Busk, J., Frost, M., Bardram, J.E., Vinberg, M. and Kessing, L.V. (2019). Objective smartphone data as a potential diagnostic marker of bipolar disorder, *Australian & New Zealand Journal of Psychiatry* **53**(2): 119–128, PMID: 30387368.

Faurholt-Jepsen, M., Vinberg, M., Debel, S., Bardram, J.E. and Kessing, L.V. (2016). Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder, *International Journal of Methods in Psychiatric Research* **25**(4): 309–323.

Gomes, H.M., Grzenda, M., Mello, R., Read, J., Le Nguyen, M.H. and Bifet, A. (2022). A survey on semi-supervised learning for delayed partially labelled data streams, *ACM Computing Surveys* **55**(4): 1–42.

González-Almagro, G., Peralta, D., De Poorter, E., Cano, J.-R. and García, S. (2023). Semi-supervised constrained clustering: An in-depth overview, ranked taxonomy and future research directions, *arXiv:* 2303.00522.

Grande, I., Berk, M., Birmaher, B. and Vieta, E. (2016). Bipolar disorder, *The Lancet* **387**(10027): 1561–1572.

Grünerbl, A., Muaremi, A. and Osmani, V. (2015). Smartphone-based recognition of states and state changes in bipolar disorder patients, *IEEE Journal of Biomedical and Health Informatics* **19**(1): 140–148.

Hryniewicz, O. and Kaczmarek-Majer, K. (2021). Possibilistic aggregation of inhomogeneous streams of data, *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Luxembourg*, pp. 1–6, DOI: 10.1109/FUZZ45933.2021.9494583.

Kaczmarek-Majer, K., Casalino, G., Castellano, G., Dominiak, M., Hryniewicz, O., Kamińska, O., Vessio, G. and Díaz-Rodríguez, N. (2022a). Plenary: Explaining black-box models in natural language through fuzzy linguistic summaries, *Information Sciences* **614**: 374–399.

Kaczmarek-Majer, K., Casalino, G., Castellano, G., Hryniewicz, O. and Dominiak, M. (2022b). Explaining smartphone-based acoustic data in bipolar disorder: Semi-supervised fuzzy clustering and relative linguistic summaries, *Information Sciences* **588**: 174–195.

Kaczmarek-Majer, K., Casalino, G., Castellano, G., Leite, D. and Hryniewicz, O. (2022c). Fuzzy linguistic summaries for explaining online semi-supervised learning, *2022 IEEE 11th International Conference on Intelligent Systems, Warsaw, Poland*, pp. 1–8.

Kamińska, O., Kaczmarek-Majer, K., Opara, K., Jakuczun, W., Dominiak, M., Antosik-Wójcińska, A., Święcicki, Ł.

and Hryniewicz, O. (2019). Self-organizing maps using acoustic features for prediction of state change in bipolar disorder, *in* M. Marcos *et al.* (Eds), *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems*, Springer, Berlin/Heidelberg, pp. 148–160.

Kamińska, O., Kaczmarek-Majer, K. and Hryniewicz, O. (2020). Acoustic feature selection with fuzzy clustering, self organizing maps and psychiatric assessments, *Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2020, Lisbon, Portugal*, pp. 342–355.

Kmita, K., Casalino, G., Castellano, G., Hryniewicz, O. and Kaczmarek-Majer, K. (2022). Confidence path regularization for handling label uncertainty in semi-supervised learning: Use case in bipolar disorder monitoring, *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Padua, Italy*, pp. 1–8.

Kusy, M. and Zajdel, R. (2021). A weighted wrapper approach to feature selection, *International Journal of Applied Mathematics and Computer Science* **31**(4): 685–696, DOI: 10.34768/amcs-2021-0047.

Lai, D.T.C. and Garibaldi, J.M. (2011). A comparison of distance-based semi-supervised fuzzy c-means clustering algorithms, *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), Taipei, Taiwan*, pp. 1580–1586.

Leite, D., Decker, L., Santana, M. and Souza, P. (2020). EGFC: Evolving Gaussian fuzzy classifier from never-ending semi-supervised data streams—With application to power quality disturbance detection and classification, *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK*, pp. 1–9.

Li, K., Cao, Z., Cao, L. and Zhao, R. (2009a). A novel semi-supervised fuzzy c-means clustering method, *Chinese Control and Decision Conference, Guilin, China*, pp. 3761–3765.

Li, Y.-F., Kwok, J.T. and Zhou, Z.-H. (2009b). Semi-supervised learning using label mean, *Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Canada*, pp. 633–640.

Low, D., Bentley, K. and Ghosh, S.K. (2020). Automated assessment of psychiatric disorders using speech: A systematic review, *Laryngoscope Investigative Otolaryngology* **31**5(1): 96–116.

Mai, D.S. and Ngo, L.T. (2015). Semi-supervised fuzzy c-means clustering for change detection from multispectral satellite image, *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Istanbul, Turkey*, pp. 1–8.

Otsu, N. (1979). A threshold selection method from gray-level histograms, *IEEE Transactions on Systems, Man, and Cybernetics* **9**(1): 62–66.

Panek, D., Skalski, A., Gajda, J. and Tadeusiewicz, R. (2015). Acoustic analysis assessment in speech pathology detection, *International Journal of Applied Mathematics and Computer Science* **25**(3): 631–643, DOI: 10.1515/amcs-2015-0046.

Pedrycz, W. and Waletzky, J. (1997). Fuzzy clustering with partial supervision., *IEEE Transactions on Systems, Man and Cybernetics B: Cybernetics* **27**(5): 787–95.

Ruiz, D. and Finke, J. (2019). Lyapunov-based anomaly detection in preferential attachment networks, *International Journal of Applied Mathematics and Computer Science* **29**(2): 363–373, DOI: 10.2478/amcs-2019-0027.

Vapnik, V. (2006). *Estimation of Dependences Based on Empirical Data*, Springer Berlin/Heidelberg.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods, *33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, USA*, pp. 189–196.

Zhou, D., Bousquet, O., Lal, T., Weston, J. and Schölkopf, B. (2003). Learning with local and global consistency, *in* S. Thrun *et al.* (Eds), *Advances in Neural Information Processing Systems*, MIT Press, Cambridge.

Zhu, X. and Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation, *Report CMU-CALD-02-107*, Carnegie Mellon University, Pittsburgh.

Gabriella Casalino is an assistant professor at the CILab—a laboratory of the Department of Informatics, University of Bari. Her research activity is focused on computational intelligence, with a particular interest for data analysis. Three are the main topics she is currently working on: intelligent data analysis, computational intelligence for eHealth, and data stream mining. She is an associate editor of the *Journal of Intelligent and Fuzzy Systems*.

Giovanna Castellano is an associate professor at the Department of Computer Science, University of Bari Aldo Moro, Italy, where she is the coordinator of the Computational Intelligence Lab. She is member of the IEEE Computational Intelligence Society, the EUSFLAT Society and the INDAM-GNCS Society. Her research interests are in the area of computational intelligence and computer vision. She has published more than 200 papers in international journals and conferences. She is an associate editor of several international journals.

Olgierd Hryniewicz is a professor at the Systems Research Institute of the Polish Academy of Sciences in Warsaw. He received his PhD and DSc degrees in 1976 and 1985, respectively. His main area of research is statistical data analysis and data mining for uncertain and imprecise data. The main area of applications of his results are quality control and reliability. He has published more than 200 papers in journals, books and conference proceedings.

Daniel Leite is an associate professor at the Department of Engineering and Science, Universidad Adolfo Ibanez, Santiago, Chile. For 9 years he had been a professor at the UFLA, UFMG and PUC-MG, Brazil. He holds a PhD from UNICAMP, Sao Paulo (2012), and was a postdoctoral researcher at the University of Ljubljana, Slovenia, in 2018–2019, and at UFMG in 2013–2014. He is a recipient of the North American Fuzzy Information Processing Society NAFIPS Early Career Award (2017), as well as PhD thesis awards from the IEEE Computational Intelligence Society (2017), NAFIPS (2015), and the Brazilian Computer Society (2014). He contributes as an associate editor of the *Evolving Systems* journal. His interests are in granular computing, control systems, and machine learning.

Karol Opara is an associate professor in the Systems Research Institute, Polish Academy of Sciences. He conducts research in the theory of metaheuristics and applied data science. His works concern various disciplines, from medicine and ecology to transportation engineering and stylometric analysis of poetry.

Weronika Radziszewska is an assistant professor in the Department of Computer Modelling in the Systems Research Institute of the Polish Academy of the Sciences, Warsaw. Her main areas of research are data processing, data analysis, multi-agent systems and control systems. She actively works in the fields of multi-energy systems, microgrids and mitigation/adaptation technologies for climate change.

Katarzyna Kaczmarek-Majer received her MS in mathematics and her MS in computer science from Adam Mickiewicz University in Poznan, Poland. Then she obtained her PhD with distinction in computer science in 2015 from the Systems Research Institute of the Polish Academy of Sciences. She is now an assistant professor at that institute. Her areas of expertise include soft computing, time series/data streams analysis and human-centered AI. She combines effectively her theoretical research with involvement in scientific projects with applications mainly in medicine and healthcare. She has co-authored 40+ scientific publications. Some of them have been awarded at scientific conferences, e.g., with the Best Paper Award at *FUZZ-IEEE 2022*. She is also the President of the Information Technologies for Psychiatry Foundation and the coordinator of the eHealth Section of the Polish Information Processing Society.

Received: 21 November 2022
Revised: 13 April 2023
Accepted: 4 May 2023